Draft version manuscript No.
(will be inserted by the editor)

Analysis of Nutrition Data by means of a Matrix Factorization Method

Jorge Díez · Edna Gamboa · Teresita González de Cossío · Oscar Luaces · Thorsten Joachims · Antonio Bahamonde

Abstract We present a factorization framework to analyze the data of a regression learning task with two peculiarities. First, inputs can be split into two parts that represent semantically significant entities. Second, the performance of regressors is very low. The basic idea of the approach presented here is to try to learn the ordering relations of the target variable instead of its exact value. Each part of the input is mapped into a common Euclidean space in such a way that the distance in the common space is the representation of the interaction of both parts of the input. The factorization approach obtains reliable models from which it is possible to compute a ranking of the features according to their responsibility in the variation of the target variable. Additionally, the Euclidean representation of data provides a visualization where metric properties have a clear semantics. We illustrate the approach with a

The research reported here is supported in part under grant TIN2011-23558 from the MICINN (Ministerio de Ciencia e Innovación, Spain). Edna Gamboa was supported by a PhD grant from CONACYT (*Consejo Nacional de Ciencia y Tecnología, México*). The paper was written while Antonio Bahamonde was visiting Cornell University with grants of *Movilidad Campus de Excelencia Internacional* (Universidad de Oviedo) and of *Programa Nacional de Movilidad de Recursos Humanos del Plan Nacional de Investigación* (Ministerio de Educación, Cultura y Deporte, Spain). The dataset was gathered in a project supported by *Ministerio de Desarrollo Social de México*

J. Díez, O. Luaces and A. Bahamonde
Universidad de Oviedo, Artificial Intelligence Center, Gijón, Asturias, Spain
E-mail: {jdiez,oluaces,abahamonde}@uniovi.es

E. Gamboa and T. G. de Cossío
Center for Nutrition and Health Research, National Institute of Public Health, Cuernavaca, Morelos, México
E-mail: magalygamboa@yahoo.com, tgonzale@insp.mx

T. Joachims
Cornell University, Department of Computer Science, Ithaca, NY, USA
E-mail: tj@cs.cornell.edu

case study: the analysis of a dataset about the variations of Body Mass Index for Age of children after a Food Aid Program deployed in poor rural communities in Southern México. In this case the two parts of inputs are the vectorial representation of children and their diets. In addition to discover latent information, the mapping of inputs allows us to visualize children and diets in a common metric space.

Keywords Matrix factorization · Learning to rank · Feature selection · Data Analysis · Nutrition data · Body Mass Index (BMI)

1 Introduction

In this paper we study the complex interactions of a set of features with a continuous target variable in a learning task. We will deal with inputs that can be split into two parts, each with its own semantics. We try to model both the interactions of these parts, and the interactions between the components of each part.

The case study that we will use throughout the paper is a data set about the increase of Body Mass Index for Age (BMI) in children after the deployment of a Food Aid Program devised to help very poor rural communities in Southern México. In this case, the parts of the input describe the children and the intervention of the Food Program, mainly children's diets.

From a formal point of view, we have a regression learning task where regressors perform very poorly in terms of mean absolute error when compared with predicting the mean value for the target variable. In the case study, the state-of-the-art regression methods only are able to reach a correlation between predictions and true values of 0.48, see details in Section 4. However, we need to find a reliable way to model the relationship between inputs defined by their features and the target. First, we want to use the model to find a ranking of the features according to their relevancy to explain the variations of the target. Additionally, we want to build graphical representations to depict the interactions with respect to the target value.

To learn a model that captures the relationships between inputs and the target variable, we will transform the original regression problem into a ranking task. Instead of trying to predict the exact target value, we will capture its ordering properties. Thus, the aim is to learn a *variation* function g that induces the same ordering on inputs than the target value.

We set an optimization problem to learn g. The idea is to learn a linear *mapping* of the objects involved in the interaction (children and diets) in a common Euclidean space $\mathbb{R}^k$. The purpose is to unveil new latent relationships from the training data. Then, g will be defined by the distance of the representations of children and diets in the common Euclidean space. The optimization problem will search for the mappings that maximize the probability of coincidence between the relative ordering induced by g and by the target values.

Roughly speaking, g will be given by a weighted sum of products of the variables that describe the children and the diets. The approach presented here *factorizes* the vector of weights in the product (in some sense defined below) of two matrices: those that define the mappings. In addition to discover latent information, this trick allows us to draw children and diets in a metric space where similarity functions can be straightforwardly defined. It is noteworthy that using this similarity, we may obtain clusters of objects according to their behavior with respect to the target value.

In other words, we will use a *matrix factorization* approach. Notice that this approach has been successfully used for a variety of application fields including, for instance, recommender systems [7], information retrieval [21,22] and construction of music playlists [14,2]. On the other hand, the factorization algorithm presented here can be seen as a visualization method to arrange in a common metric space the components of inputs in a sense that interactions are proportional to distances.

It is important to emphasize that the method proposed here is scalable to big data, notice that factorization algorithms were used in this context, see Section 2.2 for some references. Thus, we think that this paper opens some interesting possibilities to explore more subtle association in biomedical data, with the advent of new types of data and the availability of extensively linked data.

The rest of the paper is organized as follows. In the next section we give a brief description of the Food Program that produced the data analyzed as case study. Then we review the previous work related to the techniques used throughout the paper. Sections 3.1 and 3.2 present the formal setting and the learning algorithm. The ranking of features involved in the learned functional relation of inputs and outputs is built using an analysis of *sensitivity* introduced in Section 3.3. Finally, Section 4 reports an exhaustive experimentation.

2 Background

This section is devoted to firstly introduce the Food Aid Program developed in México, from which we took the data, in order to analyze which factors could determine the increase of BMI in the beneficiaries of the aid. In the second part of the section we also present some previous works in the field of machine learning which are related to the approach presented in this paper, i.e. matrix factorization and learning from partial orderings.

2.1 Food Aid Program

In the past decade, a number of programs have been launched by Governments in developing countries as an important strategy implemented to break the intergenerational transmission of poverty [10,9]. The objective of these programs is to improve the quality of life of people through interventions in health, nutrition, and education.

In 2003 Mexican Government launched a food support program called *Programa de Apoyo Alimentario* (PAL). The program provided poor households living in México in remote rural communities with either cash or a food basket (in-kind) transfers. The cash transfer was set down as the cost to the government of the food basket, the equivalent of $14.00/month in 2003. To receive the benefits of the program, a group of families had to attend nutrition and health education sessions, as well as participating in program related logistic activities.

In order to study the impact of the program, a random sample of 206 rural communities in Southern México was randomly assigned to 1 of 4 groups according to the benefits received by PAL: a monthly food basket with or without health and nutrition education, a cash transfer with education, and the control group that did not receive any benefit. Let us notice that control communities were put on the waiting list for later incorporation into the program. See [10, 9] for more details.

A number of features where registered at the beginning of the intervention (in the following, *baseline*) in 2003, and after 14 months of intervention (*follow-up*), in 2005; see [3].

The dataset used in this paper is a subset of the data so collected and was built as follows. The whole set of features was divided in two groups. The first one describes the children and their environment, especially features related to their mothers. This package has 52 features and include attributes of children at baseline as height, weight, BMI, several socio-economic indicators, and a binary feature to record whether or not the child is indigenous.

The second group of 84 features gathers the information about the group of intervention, and many features of the diet that the children had at the baseline and at the follow-up. Interventions are mainly related to diets, for this reason we called this block of features *diets*. However, the intervention may include talks given to the mothers of children about different aspects of nutrition. All features of this block are characteristics of the intervention that can be planed in advance, as fat or carbohydrates intake in children's diets.

In this paper we focus on the increase of BMI of children from the baseline to the follow-up. Thus, we excluded those features that can only be measured at the end of the follow-up and can be considered as part of the outputs of the intervention; this is the case of weight and height at follow-up.

In general, this type of support programs have been shown to have positive impacts on poverty reduction, health, nutrition, and education [11, 12, 19]. However, special care should be taken with the increase of BMI. There is an increasing prevalence of overweight in beneficiary households. For instance, the prevalence of overweight or obesity in adult women was 63.2% in the beneficiary population of PAL. The large increase in household energy consumption should thus be considered as negative for these women; see [9].

2.2 Related work

The algorithm presented in this paper is focused on the relative ordering of a variable that depends on the interactions of many others. The purpose is to provide a visualization method and to allow a feature selection of the components of the data.

The ordering aspects of the proposal of this paper are closely related to recommender systems whose aim is to present an order list of options. Thus, we used factorization models, since they constitute a very successful approach for recommender systems. For instance, the best performing algorithms in the Netflix Challenge used matrix factorization. In [7], the winners of the Challenge present this approach for recommender systems. In this case, each object to be represented in a common Euclidean space has only an identity instead of a fully vectorial description with feature values. Moreover, the interaction of objects is modeled using basically the inner product.

In [16], the authors present a general framework to learn matrix factorizations, libFM. The model presented in Section 3.2 is a bit different since the closeness variation is not included in libFM. But the general idea, presented also in previous papers [17,18] is also based on the logistic sigmoid and maximum likelihood estimation solved using *Stochastic Gradient Descend* (SGD). Other optimization methods can be used in libFM, but SGD is the most recommendable according to the authors.

Another successful factorization approach was presented in [21,22], in this case the application field is information retrieval. The aim was to learn a multilabel classifier. The authors estimate the relevancy of each label by the inner product of the representations in a common Euclidean space of the mapping of the input and each label.

More general than factorization systems are embeddings. The idea is to map objects in an Euclidean space in such a way that some function is optimized for different purposes. A use of embedding related to ordering, but in a different sense, appears in [14,2]. These papers present a probabilistic model for generating coherent *playlists* by embedding songs and social tags in a unified Euclidean space.

In [8], the authors present embeddings using kernels, and the purpose is to obtain fast retrieval methods for large-scale storage of images. In this case the Euclidean space is a low dimension Hamming space where items can be efficiently searched.

On the other hand, in [1,13,5,4], the sensitivity analysis of [15] was used to compute a ranking of features in a closely related context: learning preferences of users about a kind of items. Notice that its aim is also to learn to order things, and pairwise comparisons were also used to approach these learning tasks.

Finally, in Nutrition, the typically used tools to analyze these datasets can be divided in two blocks: Some statistical tests to check significant differences in groups, and regression methods with polynomials of degree 2, like the so-

called *double difference* that is commonly used in impact evaluation studies [9,10].

3 A proposed learning procedure

In this section we detail our proposal to analyze the data acquired from the Food Aid Program mentioned in Section 2.1. We divided this section into three parts: the first one makes an introduction to the general framework used in our proposed method, which is based in a maximum likelihood estimation by means of Stochastic Gradient Descent, thus minimizing a specific loss function explained in the section. In the general approach we introduce a function g that relates children and diets, which is explained in the second part of the section. Finally, in the third part we present a backward-chaining approach to analyze the relevance of the features representing the children/diets in the input data.

3.1 Formal Framework

Let us consider the following dataset

$$D = \{(\boldsymbol{x}_1, f(\boldsymbol{x}_1)), \ldots, (\boldsymbol{x}_n, f(\boldsymbol{x}_n))\}. \tag{1}$$

Here we assume that f is an unknown real function on a vectorial space from where inputs $\boldsymbol{x}$ are drawn.

The aim is to find a new function g of input data $\boldsymbol{x}$, that depends also on some parameters θ, such that the variations of f can be predicted by the variations of g. The function g will have an analytical definition that makes easy to compute on any input. In symbols, the aim of g is to maximize the probability

$$\Pr\big(f(\boldsymbol{x}) > f(\boldsymbol{x'}) \iff g(\boldsymbol{x}, \theta) > g(\boldsymbol{x'}, \theta)\big). \tag{2}$$

In the following we will call g the *variation* function.

To learn g, from the dataset D, we define the following ordering version

$$D_{or} = \{\big(\boldsymbol{x}_i, \boldsymbol{x}_j; [\![f(\boldsymbol{x}_i) > f(\boldsymbol{x}_j)]\!]\big) : i, j = 1, \ldots, n\}. \tag{3}$$

The symbol $[\![p]\!]$ stands for the value 1 when the predicate p is true, and -1 otherwise. In the remainder of the paper we describe an algorithm to learn g from this binary classification task.

Formally, the learning process of the parameters θ of g starts with the dataset D_{or} (3). Soon we shall see that we may use only the examples of the *positive* class,

$$D^{+}_{or} = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) : f(\boldsymbol{x}_i) > f(\boldsymbol{x}_j),\ i, j = 1, \ldots, n\}. \tag{4}$$

As usual, we assume that all these examples are independently and identically drawn (i.i.d.) from an unknown distribution. Thus, using a *maximum likelihood* approach, the parameters θ should maximize

$$L = \prod_{(i,j)\in D^+_{or}} \Pr\big(g(\boldsymbol{x}_i,\theta) > g(\boldsymbol{x}_j,\theta)|\theta\big). \tag{5}$$

We will consider the *logistic sigmoid* to represent these probabilities [17, 18, 16].

$$\Pr\big(g(\boldsymbol{x}_i,\theta) > g(\boldsymbol{x}_j,\theta)\big) = \sigma_\lambda\big(g(\boldsymbol{x}_i,\theta) - g(\boldsymbol{x}_j,\theta)\big) \tag{6}$$
$$\sigma_\lambda(x) = \frac{1}{1+e^{-x/\lambda}}.$$

In these equations, $\lambda > 0$ is a parameter used here to stabilize the numerical computations of the learning algorithm described below. Notice that if we had used the negative cases of D_{or}, we would have duplicates of each positive example, since σ_λ has the following symmetric property

$$\sigma_\lambda(-x) = 1 - \sigma_\lambda(x).$$

Following [17], the maximum likelihood estimation (5) can be done using a *SGD (Stochastic Gradient Descent)* algorithm [20] with a regularization term to ensure small components of the parameter θ. Thus, the optimal value, θ^* is given by

$$\begin{aligned}\theta^* &= \operatorname*{argmax}_\theta \ \log(L) - \nu r(\theta) \qquad (7)\\ &= \operatorname*{argmax}_\theta \sum_{(i,j)\in D^+_{or}} \log\big(\sigma_\lambda(g(\boldsymbol{x_i},\theta) - g(\boldsymbol{x_j},\theta)\big) - \nu r(\theta)\\ &= \operatorname*{argmax}_\theta \sum_{(i,j)\in D^+_{or}} -\log\Big(1+exp\big(\frac{g(\boldsymbol{x_j},\theta)-g(\boldsymbol{x_i},\theta)}{\lambda}\big)\Big) - \nu r(\theta)\end{aligned}$$

Therefore,

$$\theta^* = \operatorname*{argmin}_\theta \sum_{(i,j)\in D^+_{or}} \tilde{L}_{ij} + \nu r(\theta), \tag{8}$$

where

$$\tilde{L}_{i,j} = \log\Big(1 + exp\big(\frac{g(\boldsymbol{x_j},\theta) - g(\boldsymbol{x_i},\theta)}{\lambda}\big)\Big). \tag{9}$$

Algorithm 1 implements this approach. The algorithm starts with a random value for the parameters θ. The parameters are updated picking up a random training example from D^+_{or} and using

$$\theta \leftarrow \theta - \gamma\Big[\frac{\partial \tilde{L}_{ij}}{\partial\theta} + \nu\frac{\partial r(\theta)}{\partial\theta}\Big]. \tag{10}$$

Algorithm 1 SGD algorithm to learn the parameter θ using a Gaussian regularizer

Input: D_{or}^{+}; { (4)}
Input: $\gamma > 0$ {learning rate}; $\lambda > 0, \nu > 0$ {regularization parameters}; {r is the regularization function}:
Assign random values to the components of θ;
repeat
  fetch random $(\boldsymbol{x}_i, \boldsymbol{x}_j) \in D_{or}^{+}$;
  $\theta \leftarrow \theta - \gamma\Big[\frac{\partial \tilde{L}_{ij}}{\partial \theta} + \nu \frac{\partial r(\theta)}{\partial \theta}\Big]$;
until stopping criterion

In this equation, γ is the so-called learning rate, and ν is the regularization parameter. Both parameters must be positive. On the other hand, the partial derivatives depend on the actual definition of the variation function g.

$$\begin{aligned}\frac{\partial \tilde{L}_{ij}}{\partial \theta} &= \frac{1}{\lambda} \frac{exp\big(\frac{g(\boldsymbol{x_j},\theta)-g(\boldsymbol{x_i},\theta)}{\lambda}\big)}{1+exp\big(\frac{g(\boldsymbol{x_j},\theta)-g(\boldsymbol{x_i},\theta)}{\lambda}\big)} \Big(\frac{\partial g(\boldsymbol{x_j},\theta)}{\partial \theta} - \frac{\partial g(\boldsymbol{x_i},\theta)}{\partial \theta}\Big) \\ &= \frac{1}{\lambda}\sigma_\lambda\big(g(\boldsymbol{x_j},\theta) - g(\boldsymbol{x_i},\theta)\big)\Big(\frac{\partial g(\boldsymbol{x_j},\theta)}{\partial \theta} - \frac{\partial g(\boldsymbol{x_i},\theta)}{\partial \theta}\Big).\end{aligned} \tag{11}$$

In this section inputs were described by a generic vector $\boldsymbol{x}$ and the aim was to emphasize the ordering of these vectors according to f values. In the next section we will get into the structure of inputs as the concatenation of two different vectors, the representation of children and diets.

3.2 Mapping of Children and Diets via Matrix Factorization

In the following, we will assume that each input data can be split into two parts:

$$\boldsymbol{x} = (\boldsymbol{c}, \boldsymbol{d}).$$

We will consider an embedding of both children and diets in a common Euclidean space. Then, the function g (2) will be defined in terms of the mappings in the common space.

We assume that *children* are described by vectors in an Euclidean input space of dimension $|Ch|$, while *diets* are given by vectors with $|Di|$ components. We shall represent them in a common space of dimension k using two linear maps given respectively by matrices $\boldsymbol{W}$ and $\boldsymbol{V}$.

$$f_W : \mathbb{R}^{|Ch|} \longrightarrow \mathbb{R}^k,\ f_W(\boldsymbol{c}) = \boldsymbol{W}\boldsymbol{c} \tag{12}$$

$$f_V : \mathbb{R}^{|Di|} \longrightarrow \mathbb{R}^k,\ f_V(\boldsymbol{d}) = \boldsymbol{V}\boldsymbol{d} \tag{13}$$

Let us remark that, as usual, we are considering vectors as column matrices.

In this context, the parameters θ to be learned are the matrices $\boldsymbol{W}$, $\boldsymbol{V}$. There are different options to define the interaction of children and diets. In

this paper we present a g function that defines the interaction by the *closeness*. In symbols, we define

$$\begin{aligned} g(\boldsymbol{c},\boldsymbol{d}) &= -\left\|\boldsymbol{W}\boldsymbol{c}-\boldsymbol{V}\boldsymbol{d}\right\|^2 \\ &= -\left\|\boldsymbol{W}\boldsymbol{c}\right\|^2 - \left\|\boldsymbol{V}\boldsymbol{d}\right\|^2 + 2\langle \boldsymbol{W}\boldsymbol{c}, \boldsymbol{V}\boldsymbol{d}\rangle \end{aligned} \tag{14}$$

To allow the representation of more interactions of input data, we add one constant component (with value 1 for instance) to the vectorial representation of children and diets; that is,

$$\boldsymbol{c}^T \leftarrow [\boldsymbol{c}^T\ 1]; \quad \boldsymbol{d}^T \leftarrow [\boldsymbol{d}^T\ 1]. \tag{15}$$

Then, the variation function g includes the weighted sum of all monomials of degree 2 formed with variables taken from the description of children ($\boldsymbol{c}$) or diets ($\boldsymbol{d}$).

The derivatives needed to implement the learning algorithm are the following.

$$\begin{aligned} \frac{\partial g(\boldsymbol{c},\boldsymbol{d})}{\partial \boldsymbol{W}} &= -\boldsymbol{W}(2\boldsymbol{c}\boldsymbol{c}^T) + 2\boldsymbol{V}\boldsymbol{d}\boldsymbol{c}^T \\ \frac{\partial g(\boldsymbol{c},\boldsymbol{d})}{\partial \boldsymbol{V}} &= -\boldsymbol{V}(2\boldsymbol{d}\boldsymbol{d}^T) + 2\boldsymbol{W}\boldsymbol{c}\boldsymbol{d}^T \end{aligned} \tag{16}$$

Additionally, we shall use the square of the Frobenius norm as the matrix regularization (7).

$$r(\boldsymbol{W}) = \left\|\boldsymbol{W}\right\|_F^2 = Tr(\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{W}). \tag{17}$$

For $\boldsymbol{V}$ we use the same regularization. Therefore, the regularization derivatives are

$$\frac{\partial Tr(\boldsymbol{W}^T\boldsymbol{W})}{\partial \boldsymbol{W}} = 2\boldsymbol{W}, \qquad \frac{\partial Tr(\boldsymbol{V}^T\boldsymbol{V})}{\partial \boldsymbol{V}} = 2\boldsymbol{V}. \tag{18}$$

In the next section we describe a procedure to carry out a feature selection based on the sensitivity of the likelihood with respect to each feature.

3.3 Feature Selection

The relevance of each feature can be estimated by its *sensitivity* in a sense that will be established below. The feature ranking algorithm that we will use is a backward-chaining procedure. Following the style of the RFE (Recursive Feature Elimination) ranker [6], the feature with the lowest absolute value of the ranking criterion R

$$\underset{r,s}{\operatorname{argmin}}\ \min\{|R(r)|, |R(s)|\}, (r=1,\ldots,|Ch|; s=1,\ldots,|Di|)$$

is removed in each step. The learner is trained again with the remaining features, and the process continues until all features but one are removed. In this equation r is an index for the components of the description of the children, and s for the descriptions of the diets. A chunk of features can also be removed instead of only one at each iteration, as suggested in [6].

Notice that in this way, we obtain a ranking of the original features, and a sequence of models.

The first-order sensitivity analysis uses the derivatives of the function employed as representative of the learning process. In [15] the author introduces a virtual *scaling factor* to compute the gradient for nonlinear equations. It acts as a component-wise multiplicative term υ (whose values are 1) on the feature values.

We shall consider the sensitivity of the likelihood (7) with respect to the features. Let us recall that the likelihood is what it is optimized to learn the parameters of the model used to explain the increase of BMI. In fact, the likelihood is a function of the training set (4), D^+_{or}, and the parameters θ.

$$\mathbb{F}(D^+_{or},\theta) = \sum_{(i,j)\in D^+_{or}} \log\big(1+exp(\frac{g(\boldsymbol{\upsilon}\cdot\boldsymbol{x_j})-g(\boldsymbol{\upsilon}\cdot\boldsymbol{x_i})}{\lambda})\big). \tag{19}$$

Therefore,

$$\begin{aligned} R(r) &= \frac{\partial}{\partial \upsilon_r} \sum_{(i,j)\in D^+_{or}} \log\big(1+exp(\frac{g(\boldsymbol{\upsilon}\cdot\boldsymbol{x_j})-g(\boldsymbol{\upsilon}\cdot\boldsymbol{x_i})}{\lambda})\big) \\ &= \sum_{(i,j)\in D^+_{or}} \frac{1}{\lambda}\sigma_\lambda\big(g(\boldsymbol{x_j})-g(\boldsymbol{x_i})\big)\Big(\frac{\partial g(\boldsymbol{\upsilon}\cdot\boldsymbol{x_j})}{\partial \upsilon_r} - \frac{\partial g(\boldsymbol{\upsilon}\cdot\boldsymbol{x_i})}{\partial \upsilon_r}\Big) \end{aligned} \tag{20}$$

To implement this ranking criterion, we only need to compute the derivatives of the variation function defined in Section 3.2.

$$\frac{\partial g(\boldsymbol{\upsilon}\cdot\boldsymbol{c},\boldsymbol{d})}{\partial \upsilon_r} = 2\boldsymbol{c}_r\langle \boldsymbol{W}\boldsymbol{e^r}, \boldsymbol{V}\boldsymbol{d}-\boldsymbol{W}\boldsymbol{c}\rangle. \tag{21}$$

A similar equation can be obtained for the features of diets.

4 Experimental Results

In this section we present a report of the experiments carried out to show the benefits of the approach of this paper. We first describe the dataset used, then we present the scores achieved by the approach described in previous sections. The next subsection reports the feature rankings. The section is closed with the illustration of some graphical possibilities of the analysis of the dataset.

Table 1 Error percentages of the classifiers built with g in a train/test experiment.

k	g
2	31.13%
10	32.12%
20	33.77%
50	32.96%
100	32.28%

4.1 Datasets

The dataset used in the experiments is a subset of the data used in [3,10, 9]. We removed instances with missing values in Body Mass Index (BMI). Each example is described by 136 features; 52 are referred to *children*, and the other 84 detail the intervention done with the children. The target value is the increase of BMI from the beginning of the intervention (*baseline* in the following) to its end (*follow-up* in the following).

The imputation of missing values was done as follows. In continuous features missing values were filled with mean values computed on the training set. For discrete values (including ordinal) we used the mode.

Finally, ordinal and continuous features were standardized according to mean and standard deviation values computed on the training data.

The 2665 instances so gathered, were split in two sets: train (with 1333 elements, our D dataset), and test (with the remaining 1332). To build an ordered set of pairs, we proceeded separately in training and testing sets, and for each instance we formed 10 pairs having their target values in a different tertile. Thus, we had 13330 examples in D^{+}_{or} (4), and 13320 in test set.

4.2 Scores

Firstly, we tackle the learning task as a regression problem. For this purpose, we used Support Vector for Regression (SVR) with linear and polynomial (degree 2) kernels. The best absolute mean deviation was 0.64 achieved by a linear regressor with parameters $C = \epsilon = 0.01$. The correlation between predictions and true values was 0.48.

Moreover, if we use the trivial regressor that always predicts the mean of the target values, the absolute mean error is 0.68. That is, the best linear regressor has a 94.28% of the error committed by the trivial mean predictor.

The scores reported in Table 1 are the error percentages achieved in a simple train/test experiment. The table shows scores of the classifiers built with g (14) for different values of k (12). In all cases, we used $\lambda = 500$. Each cell of the table reports the best score achieved by the corresponding algorithm (column) with a value of k (row) when, $\gamma \in \{0.1, 0.3\}$, and $\nu \in \{0.01, 0.05\}$. In the same learning task, the error percentage of the best linear regressor was 36.2, clearly worse than the scores of Table 1 and Table 2 explained below.

Table 2 Percentages of misclassified pairs of the test set for classifiers learned (with $k = 2$ and $k = 20$) using different number of top ranked features according to the criterion of Section 3.3.

#feat	$k = 2$	$k = 20$
136	31.13%	33.77%
80	30.70%	35.25%
20	31.40%	31.49%
10	30.50%	31.22%
5	30.35%	**30.09%**
4	**30.49%**	38.40%
3	38.99%	39.03%
2	38.75%	38.64%
1	40.57%	40.57%

We did not try to find the best value for parameters using an internal grid search since the purpose was not to make a comparison of the models learned by different variation functions. In fact, we appreciate that the scores shown in Table 1 are quite similar, ranging from 31% to almost 34%. The best values are obtained for $k = 2$, which allows us to draw the mapped data in $\mathbb{R}^2$, revealing some hidden patterns in the data.

4.3 Feature Selection

Let us recall that we are trying to find the features that are more useful in order to predict the increase of BMI. This is not exactly the value of the BMI at the beginning of the process.

Table 2 shows the error rates obtained in the test set using the top ranked features with different values of k, 2 and 20. In bold fonts we highlighted the scores achieved by the best trade-off between error proportion and number of features involved. In both cases, the error rate increases dramatically if we remove only one feature more.

The best scores are achieved with 5 or 4 features depending on the value of k. These features selected for $k = 2$ are the following: z-scores of height and weight, and age of the child at baseline, and the intake of fat at the follow-up. These features reach a proportion of errors in the test set of 30.49%. If we use $k = 20$, the variables selected are the same with the addition of intake of fat at the baseline. The errors are also quite similar, in this case, 30.09%.

It is important to realize that the whole set of variables is quite redundant. In fact, if the 4 most meaningful features with $k = 2$ are removed, then the best score is obtained with a set of 30 features that achieve an error rate of 35.65%.

4.4 Graphical Representations

An important contribution of the approach presented here is the possibility of producing a graphical representation of the data from a dataset like D in (1). In this section we present two figures which illustrate this capability. In both cases we used $k = 2$ and the whole dataset D for training. The predictions that appear in the figures are computed over all available data.

In Figure 1 we depict the relationship between predicted and true values of BMI variation (f-values in dataset D). Points were aggregated in 10 bins of equal frequency for g-values, and in 20 bins for f-values. Notice that this relation can not be established in terms of exact values since regressors, in the best case, have only an accuracy similar to the trivial predictor that always returns the mean value. Thus, we had focused on the ordering relations of inputs according to f. Table 1 reports the goodness of the predictions, but Figure 1 shows the boxplot of predictions and true values for the whole dataset. To avoid dispersion, we grouped g-values in 10 bins of equal frequency (represented in the horizontal axis), and f-values in 20 bins (represented in vertical axis).

In the figure we can see that the ordering provided by g and f are coherent only to a certain degree; the estimation of the error was 31.13%. It is possible to argue that this score is not too high, but this is the best relationship that we can learn. Recall that it is a very difficult task to figure out the exact value of f using a regressor. In any case, we can clearly appreciate that median values in Figure 1 increase with g-predictions.

The graph in Figure 2, gives a visual representation of the increase of BMI in an specific and very important group of children, the indigenous that constitute the 27.7% of the data. We have drawn the mappings of the mean individuals of eight groups of children: indigenous and non-indigenous in each of the four kinds of interventions of the Food Program. The closer two points are in the picture, the more similar their increase of BMI. We can recognize that there is a big difference between these two groups of children. Let us remark that indigenous are an especially depressed ethnic in México even if they are compared to poor children of rural areas.

It is noteworthy the different role played by education as a complement to food basket. Let us recall that PAL was a conditioned program, which means that the food assistance was conditioned on attending nutrition and health education sessions, as well as participating in program-related logistic activities. However, one of the four kinds of intervention was food basket without education. Thus, it is an interesting issue to test the benefits of education in BMI increase.

The representation of groups with and without education is too near in indigenous children. This suggests that education was not important; in fact, for this group, food basket is relevant since it makes big differences with cash transfer and the control group. On the other hand, we can observe that education was important for non-indigenous since there is a big distance from the groups with food basket that can only be due to education. Moreover, notice

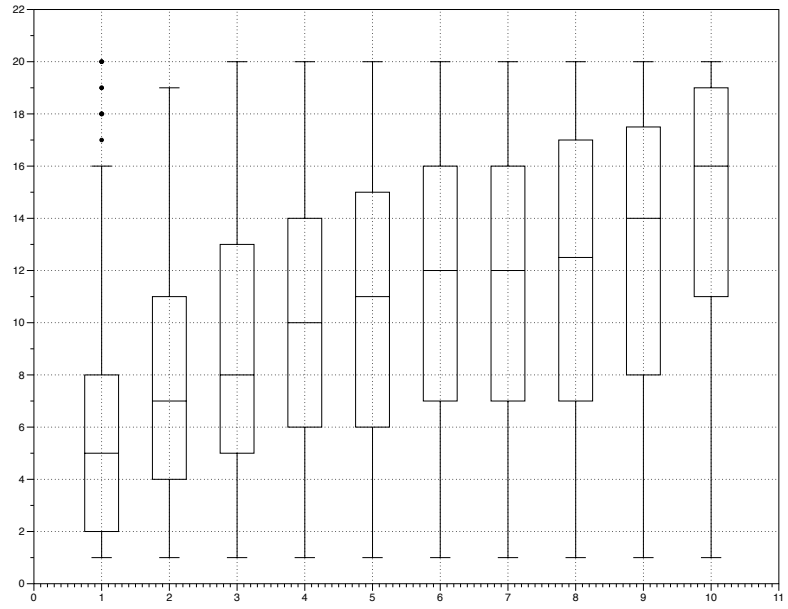

Figure 1 Boxplot of the increase of BMI predicted by g (horizontal axis) and the true variation (vertical axis) represented by f, (1).

that in non-indigenous cash transfer is closer to food basket with education, because both kinds of interventions included education.

One possible explanation for the irrelevancy of education in indigenous may be the language. There could be some problems to understand the talks about nutrition since the knowledge of Spanish is not very good in the indigenous population. Additionally, the cultural level of the mothers (typically the attendees of the talks) prevent from a good exploitation of the talks.

5 Concluding remarks

We have presented an alternative method to analyze regression learning tasks with two important characteristics: inputs are given by two vectors with their own semantics, and the performance of regressors is very low.

The approach presented in this paper learns from the relative ordering of the target variable instead of trying to predict its exact values. For this purpose, we map each part of the input into a common Euclidean space. The metric properties of that space allows us to formulate the learning task as a solvable optimization problem.

Once we have a reliable model of the relationship of inputs and target variable, we can search for a ranking of the relevancy of input features with respect to their relevancy. Another important side effect of this approach is that we

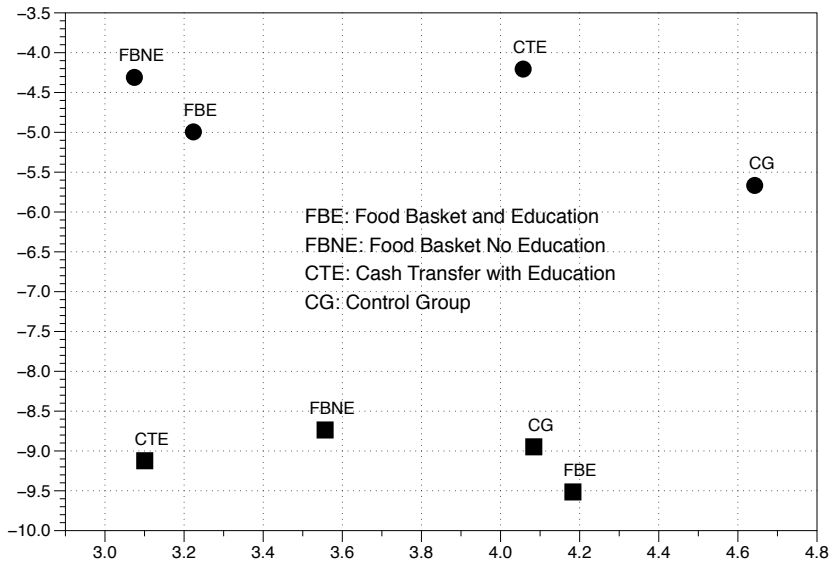

Figure 2 Centroids of indigenous (circles) and non-indigenous (squares) children in groups by their type of intervention.

can draw the objects involved in the dataset in a graphical representation with a meaningful semantics.

We present the results achieved with a dataset about the BMI increase of children in Southern México after the deployment of a Food Program. In this case inputs are split in features that describe child and features of their diets.

The approach presented let us visually inspect the results of the mapping, which allows us to gain insight into the problem from a graphical perspective. As an example, we presented a picture of the locations on the map of indigenous compared to non-indigenous children depending on the type of intervention in the Food Program. Visually, we may appreciate the big differences in terms of BMI increase of both ethnic communities.

References

1. Bahamonde, A., Bayón, G.F., Díez, J., Quevedo, J.R., Luaces, O., del Coz, J.J., Alonso, J., Goyache, F.: Feature subset selection for learning preferences: A case study. In: Proceedings of the International Conference on Machine Learning (ICML '04), pp. 49–56 (2004)
2. Chen, S., Moore, J., Turnbull, D., Joachims, T.: Playlist prediction via metric embedding. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 714–722. ACM (2012)
3. González de Cossío, T., Gutiérrez, J., González-Castell, D., Rodríguez-Ramárez, S., Unar, M., Leroy, J., Gadsden, P., Hernández-Licona, G., Gertler, P.: Evaluación de impacto del programa de apoyo alimentario. In: Nutrición y pobreza: política pública basada en evidencia. World Bank, SEDESOL (2008)

4. del Coz, J.J., Bayón, G.F., Díez, J., Luaces, O., Bahamonde, A., Sañudo, C.: Trait selection for assessing beef meat quality using non-linear SVM. In: Advances in Neural Information Processing Systems 17 (NIPS '04), pp. 321–328 (2005)
5. Díez, J., Bayón, G.F., Quevedo, J.R., del Coz, J.J., Luaces, O., Alonso, J., Bahamonde, A.: Discovering relevancies in very difficult regression problems: applications to sensory data analysis. In: Proceedings of the European Conference on Artificial Intelligence (ECAI '04) (2004)
6. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine learning **46**(1-3), 389–422 (2002)
7. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8), 30–37 (2009).
8. Kulis, B., Grauman, K.: Kernelized locality-sensitive hashing for scalable image search. In: Proceedings of the IEEE 12th International Conference on Computer Vision, pp. 2130–2137. IEEE (2009)
9. Leroy, J.L., Gadsden, P., González de Cossío, T., Gertler, P.: Cash and in-kind transfers lead to excess weight gain in a population of women with a high prevalence of overweight in rural mexico. The Journal of Nutrition **143**(3), 378–383 (2013)
10. Leroy, J.L., Gadsden, P., Rodríguez-Ramírez, S., Gonzalez de Cossío, T.: Cash and in-kind transfers in poor rural communities in mexico increase household fruit, vegetable, and micronutrient consumption but also lead to excess energy consumption. The Journal of Nutrition **140**(3), 612–617 (2010)
11. Leroy, J.L., García-Guerra, A., García, R., Dominguez, C., Rivera, J., Neufeld, L.M.: The oportunidades program increases the linear growth of children enrolled at young ages in urban mexico. The Journal of Nutrition **138**(4), 793–798 (2008)
12. Leroy, J.L., Ruel, M., Verhofstadt, E.: The impact of conditional cash transfer programmes on child nutrition: a review of evidence using a programme theory framework. Journal of Development Effectiveness **1**(2), 103–129 (2009)
13. Luaces, O., Bayón, G.F., Quevedo, J.R., Díez, J., del Coz, J.J., Bahamonde, A.: Analyzing sensory data using non-linear preference learning with feature subset selection. In: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD '04), pp. 286–297 (2004)
14. Moore, J., Chen, S., Joachims, T., Turnbull, D.: Learning to embed songs and tags for playlist prediction. In: Proceedings ISMIR (2012)
15. Rakotomamonjy, A.: Variable selection using svm based criteria. The Journal of Machine Learning Research **3**, 1357–1370 (2003)
16. Rendle, S.: Factorization machines with libfm. ACM Transactions on Intelligent Systems and Technology (TIST) **3**(3), 57 (2012)
17. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: Bpr: Bayesian personalized ranking from implicit feedback. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp. 452–461. AUAI Press (2009)
18. Rendle, S., Schmidt-Thieme, L.: Pairwise interaction tensor factorization for personalized tag recommendation. In: Proceedings of the third ACM international conference on Web search and data mining, pp. 81–90. ACM (2010)
19. Rivera, J.A., Sotres-Alvarez, D., Habicht, J.P., Shamah, T., Villalpando, S.: Impact of the mexican program for education, health, and nutrition (progresa) on rates of growth and anemia in infants and young children. JAMA: the journal of the American Medical Association **291**(21), 2563–2570 (2004)
20. Robbins, H., Monro, S.: A stochastic approximation method. The Annals of Mathematical Statistics pp. 400–407 (1951)
21. Weston, J., Bengio, S., Hamel, P.: Multi-tasking with joint semantic spaces for large-scale music annotation and retrieval. Journal of New Music Research **40**(4), 337–348 (2011)
22. Weston, J., Bengio, S., Usunier, N.: Large scale image annotation: Learning to rank with joint word-image embeddings. Machine learning **81**(1), 21–35 (2010)